Speaker Adaptation Using Multiple Reference Speakers

Francis Kubala, Richard Schwartz, Chris Barry

BBN Systems and Technologies Corporation
Cambridge, MA 02138

ABSTRACT

We introduce a new technique for using the speech of multiple reference speakers as a basis for speaker adaptation in large vocabulary continuous speech recognition. In contrast to other methods that use a pooled reference model, this technique normalizes the training speech from multiple reference speakers to a single common feature space *before* pooling it. The normalized and pooled speech can then be treated as if it came from a single reference speaker for training the reference hidden Markov model (HMM). Our usual probabilistic spectrum transformation can be applied to the reference HMM to model a new (target) speaker. In this paper, we describe our baseline (single reference speaker) speaker-adaptation system and give current performance results from a recent formal evaluation of the system. We also describe our proposal for adapting from multiple reference speakers and report on recent preliminary experimental results in support of the proposed technique.

1 INTRODUCTION

We have, in the past, reported our work in speaker adaptation for large vocabulary continuous speech recognition using a probabilistic spectral mapping [5]. In that work we transformed well-trained phonetic hidden Markov models of a single reference speaker so that they were appropriate for a new (target) speaker. This method reduced the recognition error rate by about a factor of five relative to a cross-speaker model (trained on one speaker, tested on another). However, the resulting error rate was still 2 to 3 times that obtained with a speaker-dependent model for the target speakers.

In recent years several researchers have demonstrated speaker-independent recognition using essentially the same recognition algorithms used for speaker-dependent recognition, but with a model derived by simply pooling the training speech of over 100 speakers as if it all were produced by one speaker. For these systems, the error rate is again 2 to 3 times that of speaker-dependent models. This shows that there is value in simple pooling of data from many speakers. The logical extension of these two results would be to use the pooled speaker-independent model as a reference model for speaker adaptation. However, we know that pooled training yields a model that has very broad (less discriminating) distributions compared to those produced by speaker-dependent training. Since the adaptation procedures that we have investigated also smooth the original model, we expect that a straightforward application of them to a pooled speaker-independent model will fail to yield improvements due to excessive smoothing.

The approach we propose here consists of three steps:
1) To reduce the smearing of the model distributions, we estimate and apply a deterministic spectral transformation to each reference speaker so that their speech parameters lie in a single common space.
2) We then treat all the transformed speech as if it came from one speaker for training the reference HMM.
3) Finally, we estimate and apply our usual probabilistic spectrum transformation to the pooled reference HMM to model a new target speaker.

In the next section, we describe our basic speaker-adaptation system in terms of its two primary speaker-transformation strategies; speech normalization and PDF mapping. Section 3 contains experimental results which establish our current performance for a single reference speaker system and introduce preliminary evidence in support of our proposal for using multiple reference speakers.

2 BASELINE SYSTEM DESCRIPTION

Our current baseline speaker-adaptation system consists of two distinct components, both of which estimate transformations between the reference and target speaker, with the goal of making one of them 'look' like the

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1989		**00-00-1989 to 00-00-1989**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Speaker Adaptation Using Multiple Reference Speakers	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **BBN Technologies,10 Moulton Street,Cambridge,MA,02238**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**7**	

other. The first component estimates a deterministic transformation which is applied to the speech features of the reference (target) speaker. After transformation within this *speech normalization* component, the speech features of the reference (target) speaker are superimposed upon the feature space of the target (reference). The second component estimates a probabilistic transformation which is applied to the HMM parameters of the reference speaker. After transformation by this *PDF mapping* component, the modified reference model can be used as an approximation to a well-trained HMM for the target speaker. These two primary components of the system are described in more detail below.

2.1 Speech Normalization

Speech normalization is accomplished by aligning the speech features of the reference and target speakers from a small training set of utterances of known (supervised) and pair-wise identical (script-dependent) transcription. Dynamic time warping (DTW) is used to derive the alignment of a given pair of utterences. The alignments can then be used to estimate a deterministic non-parametric transformation to describe differences in the feature spaces of the two speakers. Any unsupervised feature conditioning which can be applied prior to the DTW is also performed by the speech normalization component.

The normalization procedure has been described in [2] and is briefly summarized here:

1. Make a VQ codebook for one of the speakers.
2. Partition the feature space of one speaker by quantizing that speaker's training speech.
3. Map the partitioning to the other speaker through the DTW alignment.
4. Compute the means of each sub-population defined by the VQ and the mapped-VQ.
5. Shift the features of one speaker by the difference in the means of the corresponding sub-populations.
6. Go to (3) if the alignment MSE has not converged.

The speech normalization procedure is typically applied iteratively since each application of steps (3) and (5) above reduce (or leave unchanged) the MSE of the alignment. Note that, in this procedure, the codebook is used only to partition the space of one speaker into compact regions to define the degrees of freedom in the non-parametric mapping between the speakers. The alignment of the paired utterances is computed on the original (unquantized) speech features.

2.2 PDF Mapping

PDF mapping is accomplished by aligning the (normalized) speech features of the reference and target speaker, again using DTW. This final alignment serves to define a pair-wise correspondence between the VQ spectra of the reference and target speakers which can be used to estimate a probabilistic mapping between them. The VQ spectra are determined by independent codebooks made for each of the speakers. The codebook for the target speaker is made from the limited training material available for adaptation. The computed mapping is then used to modify the discrete HMM observation density parameters of the reference model.

The mapping procedure has been described in [1] and is summarized here:

1. Make VQ codebooks for both speakers.
2. Quantize the target and reference training speech.
3. Use DTW to define a set of co-occurring VQ pairs.
4. Accumulate frequency counts of the VQ co-occurrences into a count matrix.
5. Normalize the count matrix yielding a transformation matrix.
6. Apply the transformation matrix to the reference HMM (discrete) observation densities.

The resulting transformed model is then used directly in recognition as if it were a model derived from the target speaker.

The transformation described above can be made more detailed by defining a set of class-dependent matrices and labeling the states of the reference HMM with their class membership. One easily implemented set of equivalence classes for a phoneme-based system such as BYBLOS is the set of phoneme-dependent transformations defined by the phonemes in the lexicon. Since the reference speaker has provided enough speech to train a high-performance speaker-dependent HMM, the model can be used to automatically label the reference speech prior to computing the spectral mapping.

3 EXPERIMENTAL RESULTS

Speaker-adaptation encompasses a wide variety of practical scenarios. Our current speaker-adaptation algo-

rithms are suited to a batch-style, limited-training scenario appropriate for bringing a new speaker up to an acceptable initial recognition performance level. We have concentrated on the new speaker start-up scenario, using supervised techniques on a small set of known training utterances, in the belief that supervised techniques are most likely to succeed in the short term.

In most of our development work, and in the experiments described below, we have used the speaker-dependent data from the 1000-word DARPA Resource Management continuous speech database [4]. All results reported here used 2 minutes (40 utterances) of adaptation material (limited-training) from the target speaker. The standard word-pair grammar, defined as part of the database for evaluation purposes, was used in all cases except where specified otherwise. The number of test speakers and the identity of the test set vary across the experiments described below, and are noted where important. Unless otherwise noted, the performance numbers given for all experiments are:

% Word Error = 100 × [(*substitutions* + *deletions* + *insertions*) / *total number of word tokens*]

3.1 Single Reference Speakers

We performed a series of experiments on the baseline (single reference speaker) system to examine several issues related to our proposal for using multiple reference speakers. The experiments described below investigate the importance of feature conditioning for DTW, determine the effect of reference speaker identity on the estimation of the between-speaker transformations, and establish a baseline performance level for the single reference speaker system.

3.1.1 Feature Conditioning

The raw speech parameters that we use (Mel-warped cepstra, cepstral differences, normalized and difference energy) have widely varying dynamic ranges. This necessitates some form of feature pre-conditioning to avoid degenerate alignments from the DTW.

In the past, we have found that normalizing each feature independently to unit-variance (computed over the adaptation utterances) provided a satisfactory and convenient solution to the dynamic range problem. After such a normalization, each feature contributes equally on average to the alignment score computed by DTW. This simple approach performed marginally better than weighting the original feature vectors to equalize the contribution of each feature set to the DTW score.

Condition	Norm Only	+ PDF Mapping
1) Unit Variance	52.1	6.9
2) + Zero-Mean	45.1	7.6
3) + Weighting	34.2	5.7

Table 1. Improvements in speech normalization from feature conditioning.

The results shown in Table 1 compare three cases of feature conditioning, tested on six speakers. The results given in the column labeled, *Norm Only*, were achieved by computing the feature transformation from the target adaptation speech and applying it to the target's test speech. The transformed target speech was then quantized by the reference codebook and recognized using the reference (cross-speaker) HMM. The results given in the column labeled, + *PDF Mapping*, were achieved after applying the PDF spectrum tranformation to the reference HMM. For this condition, the target speech is quantized by a speaker-dependent codebook made from the target's adaptation speech. The PDF mapping is therefore computed between two independent codebooks as in our standard baseline system.

The unit variance condition (1) establishes a baseline performance for the system. This condition is similar to the system configuration used for the results from Feb. '89 reported in [3]. For condition (2) in the table, the sample mean is removed from the speech features of both reference and target speakers after normalizing the features to unit variance. This yields a small improvement for the *Norm Only* case but doesn't improve when the PDF transformation is used. Condition (3) applies a fixed, non-unit weighting to the features of both speakers after unit variance scaling and mean removal. This yields an additional 25% reduction in error for the normalization alone and marginally improves the performance of the PDF mapping. For this condition the cepstral features (unit-variance normalized) of both speakers were scaled by the square root of the cepstral index of the feature. The normalized energy feature was scaled by $\sqrt{2}$, while the difference energy was left unchanged at unit variance.

These results indicate that the DTW is sensitive to feature conditioning when computing alignments for the purpose of estimating a between-speaker normalization.

This indicates that further work is needed in feature conditioning and suggests that improvements to the iterative normalization procedure itself may also be important. It is also evident that the performance of the PDF mapping is largely independent of the quality of the normalization. This result is important since we must rely more heavily upon the feature normalization procedure when using multiple reference speakers as we propose.

3.1.2 Current Baseline Results

We tested our baseline, single reference, speaker-adaptation system on new test data for the Oct. '89 DARPA speech recognition evaluation. We used the feature conditioning enhancements described above. In addition the models for our standard reference speaker were retrained using cross-word-boundary context-dependent triphones.

The reference model was trained from 30 minutes of speech (600 utterances). Two minutes of speech (40 utterances) from the target speaker was used to compute the transformations. All development was done on the designated, May88 test set, consisting of 25 utterances per speaker.

The twelve speaker average word error rate for the Oct. '89 test set was 7.4% for the word-pair grammar and 28.7% for the no-grammar condition. These average results are competitive with the best speaker-independent results being reported today (elsewhere in these proceedings) on different, but comparable, test data. While the speaker-independent scenario requires no adaptation speech from the target speaker, it does require a large training data collection effort to provide adequate training for the (pooled) reference model. Specifically, speaker-independent training for the DARPA evaluations utilizes about 3.5 hours (4000 utterances) of reference speech from over 100 speakers. In contrast, our baseline speaker-adaptation system uses only 30 minutes (600 utterances) of reference speech from a single speaker to achieve the same performance. This suggests that speaker-adaptation may offer a more economical approach for those applications which require rapid configuration on new task domains.

Detailed Oct. '89 evaluation results for the word-pair grammar are shown in Table 2 in order of increasing word error rate. The results in the last column of the table are:

% Word Correct = $100 \times [1 - (substitutions + deletions) / total\ number\ of\ word\ tokens]$.

Speaker	Word Error	Word Correct
HXS (F)	2.0	98.3
DMS (F)	3.0	97.5
JWS	3.3	96.7
DAS (F)	3.7	97.3
DTD (F)	4.8	95.2
TAB	4.9	95.8
PGH	5.0	95.0
DTB	8.2	93.6
CMR (F)	9.9	92.1
BEF	13.3	87.2
RKM	13.8	87.7
ERS	17.2	85.3
AVG	7.4	93.5

Table 2. Baseline speaker-adaptation system results for Oct. '89 evaluation test with word-pair grammar.

Curiously, the female target speakers tend to achieve higher recognition results despite the fact that the reference speaker is male. Also, these results show a wide variance across speakers that is not consistent with speaker-dependent results (elsewhere in these proceedings) obtained from these same speakers on the same test material.

In order to prove useful, speaker-adaptation must perform reliably for most speakers, and must be considerably more powerful than can be demonstrated today. Below, we discuss several possible strategies for improving our speaker-adaptation performance.

3.1.3 Alternate Reference Speakers

In all of our previous work in speaker adaptation, one particular speaker has been used as the reference. Here we investigated the effect of the reference speaker's identity on recognition performance. Our standard speaker (male) was recorded at BBN in a normal office environment and spoke in a clear deliberate style. The development training and test data, on the other hand, was collected at another site in a sound isolating booth, and the subjects (both male and female) often spoke in casual undirected styles.

We tested the effect of reference speaker identity by selecting four additional speakers from the database to be used as reference speakers. The speakers were chosen with the sole criterion that their speaker-dependent mod-

els performed better than the average of the 12 speakers in the database.

Reference	Word Err
RS	11.8
TAB	10.7
PGH	11.7
DMS (F)	12.2
DTD (F)	14.9
Average	12.3

Table 3. Comparison of alternate reference speakers used for speaker-adaptation.

In Table 3, we show results, averaged over five test speakers, for each of five reference speakers. Speaker, RS, is our standard reference speaker. The results show that selection of an adequate reference speaker is not a difficult task since three of the four new speakers chosen do as well as our standard speaker. Furthermore, the recording and speaking style differences between our standard reference speaker and the test speakers are apparently not important ones, since reference speakers selected from the homogeneous database material did no better than our standard speaker. The 2σ confidence interval for this experiment is $\sim \pm$ 2.2%.

The alternate reference speaker results were also used to determine whether the individual pairings of reference and target speaker were important. Since each target speaker had been adapted to each of the 5 reference speakers, we could pick the best matching reference for each target based on overall recognition performance.

Target	Best Reference	Word Err
BEF	TAB	9.8
CMR	DMS	12.8
DTB	TAB	3.5
JWS	DTD or RS	9.0
RKM	TAB	14.5
Average		9.9

Table 4. Post-hoc selection of best reference speaker for a given target speaker.

The resulting average word error rate for (unfair) post-hoc reference selection was 9.9% as show in Table 4. This is 20% less than the average across all target-reference combinations shown in Table 3. This result represents an upper bound on the improvement that could be expected from automatic reference speaker selection at the test set level, making such a strategy relatively unattractive.

Since we need a larger improvement than seems likely from any single reference speaker, we are attempting to find effective methods of combining multiple reference speakers.

3.2 Multiple Reference Speakers

We have performed two preliminary experiments to explore the feasibility of combining multiple reference speakers for speaker-adaptation.

3.2.1 Averaged Reference Models

One approach for combining multiple reference speakers into a single reference model is to adapt each reference speaker independently to the target speaker, and use the adapted models jointly in the recognition stage. A straight-forward method of combining the adapted models is to average the HMM (discrete) densities.

We created such a combined reference model from the last 4 of the reference speakers shown in Table 3. The resulting recognition word error rate for the averaged model was 9.3%, compared to 12.4% for the average of the same 4 speakers used as single reference speakers. While this result is encouraging, the gain must be measured against the added expense of the scenario. Also this approach produces a more smoothed adapted model than the single reference baseline system, so that it may not extend to combinations of large numbers of reference speakers.

In order to reduce the smoothing inherent in averaging HMM parameters, we have tried combining the reference speakers before the final adapted model is trained.

3.2.2 Pooled Normalized Reference Speech

The feature normalization component of our system is designed to superimpose the speech features of one speaker onto another's for the purpose of improving the DTW alignment used for estimating the PDF mapping. This same component can be used to transform the features of many reference speakers to a single, common speaker (a *prototypical reference* speaker). The transformed speech can then be pooled and trained as if it

came from a single reference speaker. The resulting model parameters should be less smoothed (more discriminating) than a model made from similarly pooled, but unnormalized speech.

A target speaker can be similarly normalized to the prototypical reference before adapting with the PDF mapping component of the system, exactly as is done in our standard single-reference speaker-adaptation system.

Condition	X-Spkr	Norm Only	PDF Map Only	Norm + PDF Map
1 Ref	99	52.1	9	6.9
12 Ref	?	10.4	?	7.3

Table 5. Comparison of single and multiple reference systems, with speech normalization and PDF mapping.

Preliminary results from an experiment designed to test this proposal are shown in Table 5. The table compares performance for a single reference speaker against a 12 speaker reference model across four conditions:
1) cross-speaker recognition (train on reference speaker(s), test on target speaker)
2) speech normalization before cross-speaker recognition
3) PDF transformation of cross-speaker model to adapted target model
4) speech normalization before PDF transformation of cross-speaker model.

All conditions in Table 5 are based on the results from 6 target speakers on the designated May88 test set. Two minutes of speech (40 utterances) from the target speaker were used to estimate the speaker transformations. The single reference condition used 600 utterances from our standard reference speaker, RS, to train the reference model. For the 12 speaker reference condition, 11 speakers were normalized (the intended target speaker was held-out) to the prototypical reference speaker, RS. This resulted in a pool of 7200 normalized training utterances for each target speaker. A single codebook was made for the entire experiment from 100 utterances from each of the 13 speakers. The normalization used in this experiment did not include the feature conditioning improvements described earlier. The baseline unit-variance feature scaling was used here. Note that condition (1) shown in Table 1 is identical to the single reference condition, with normalization only, shown here in Table 5.

The single reference results show that normalization alone halves the error rate relative to cross-speaker recognition, while PDF mapping alone yields a ten-fold reduction in error rate. When combined, however, the additional gain is small. In the past, this effect has led us to regard the normalization as a way to make small improvements to the DTW-based alignment used for computing the PDF transformation.

The 12 speaker results, however, show that the normalization alone can be made as powerful as the PDF mapping by utilizing speech from multiple reference speakers. A five-fold reduction in error rate is realized for normalizing 12 reference speakers instead of one. Since the 12 speaker unnormalized control condition (pooled cross-speaker) has not been completed at this writing, we cannot say what proportion of the improvement is due to the normalization procedure, the additional training speech, and the additional reference speakers. As was the case for the single reference condition, combining the two transformations yields only a small additional improvement.

While these absolute performance numbers are unimpressive, pooling the normalized speech of only 12 speakers has realized a dramatic reduction in error rate over the single reference normalization. At this point, it makes sense to ask: How much better would this condition be if done on 100 reference speakers? The speaker-independent portion of the DARPA Resource Management database will permit us to answer this question.

4 Summary

We have described our speaker-adaptation system in terms of the two speaker-transformations used to make one speaker look like another; *speech normalization* and *PDF mapping*. Experimental results indicate that the speech normalization can be improved by feature conditioning, whereas the PDF mapping is relatively insensitive to improvements in the normalization. Also we have shown that the choice of any single reference speaker is not an important issue, indicating that improvements to the reference model are likely to be gained only by using multiple reference speakers.

We have reported baseline system (single reference speaker) test results of 7.4% word error rate for the word-pair grammar and 28.7% for no grammar on the designated Oct. '89 DARPA evaluation test set. This performance is comparable to the best speaker-independent results being reported today, but with considerably less

effort required to collect the reference training material (1 speaker vs. 100, 600 utterances vs. 4000).

We have proposed a new method of utilizing speech from multiple reference speakers by transforming them to a single common feature space before pooling. Preliminary experiments have shown a five-fold reduction in error rate for using the proposed normalization on a 12 speaker pooled model compared to a single speaker model. We propose to test our approach on the speaker-independent portion of the DARPA Resource Management database in the near future.

Acknowledgement

This work was supported by the Defense Advanced Research Projects Agency and monitored by the Office of Naval Research under Contract No. N00014-85-C-0279.

References

[1] Feng, M., F. Kubala, R. Schwartz, J. Makhoul (1988) "Improved Speaker Adaptation Using Text Dependent Spectral Mappings," *IEEE ICASSP-88*, paper S3.9.

[2] Feng, M., R. Schwartz, F. Kubala, J Makhoul (1989) "Iterative Normalization for Speaker-Adaptive Training in Continuous Speech Recognition," *IEEE ICASSP-89*, paper S12.4.

[3] Kubala, F., M. Feng, J. Makhoul, R. Schwartz (1989) "Speaker Adaptation from Limited Training in the BBN BYBLOS Speech Recognition System," *Proceedings of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann Publishers, Inc., pp. 100-105, Feb. 1989.

[4] Price, P., W. Fisher, J. Bernstein, and D. Pallett (1988) "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," *IEEE ICASSP-88*, paper S13.21.

[5] Schwartz, R., Y. Chow, F. Kubala (1987) "Rapid Speaker Adaptation using a Probabilistic Spectral Mapping," *IEEE ICASSP-87*, paper 15.3.1.